CHAPTER 1

INTRODUCTION

1.1 Introduction

Malaysia is rich with chemical diversity of its natural products. It is estimated, there are about 12,000 species of plants found in this country and more than 1000 species are said to have therapeutic properties [1]. Much of these resources are still untapped although a number of research groups have been actively involved in systematically studying their chemical and biological properties. Some of these compounds and their derivatives have been shown to have antibacterial properties [2, 3]. For example, bioactive compounds can be produced from the family of *Rubiaceae, Verbanaceae, Zingiberaceae and Piperaceae.*

Tuberculosis, mainly caused by *Mycobacterium tuberculosis,* is the leading killer among all infectious disease worldwide and is responsible for more than two million deaths annually. The recent increase in the number of multi-drug resistant clinical isolates of *M. tuberculosis* has created an urgent need for discovery and development of new anti tuberculosis lead compounds. It is expected that the quantitative structure-activity relationship (QSAR) approach which has been successfully applied to study factors involved in determining chemical properties or biological activities of chemical compounds can be applied here [4].

In a typical structure-activity relationship study, one is interested to develop models that can correlate the structural features of a series of chemical compounds

with their physicochemical properties or biological activities. These correlation models can be used to predict the activity of new compounds as well as to form a basis for understanding factors affecting their activities [5, 6].

QSAR models are constructed by analyzing known or computed property data and series of numerical descriptors representing the structural characteristic. Descriptors quantitative properties depend on the structure of the molecule. Various physicochemical parameters including thermodynamic properties (such as system energies), electronic properties (e.g. value of highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), molecular shape (e.g. surface area, length to breadth ratio) and simple structural characteristic (e.g. number of bonds, connectivity indexes, etc) have been used to get solid models which were able to predict the biological activity of unknown molecules [6].

In this study the structure activity relationship approach above was implemented to develop models that can correlate structural features of the compounds isolated from plants with their anti bacterial activity. Good models developed using the method were applied to screen a large chemical database. Results of the screening probes can be used to select and to postulate structure of leads molecules that can be synthesized in the production of new drugs in pharmaceutical industries.

1.2 Quantitative Structure Activity Relationship (QSAR)

Drugs exert their biological effects by participating in a series of events which include transport, binding with the receptor and metabolism to an inactive species. Since the interaction mechanism between the molecule and the putative receptor are unknown in most cases (i.e., no bound crystal structure), one is reduced to making inferences from properties which can easily be obtained (molecular properties and descriptors) to explain these interactions for unknown molecules.

The pharmaceutical companies need to continuously discover and develop new drugs, particularly in the field of anti-infective agents, in order to fight the increase of resistance to older drugs and newly discovered types of infections such as mutated bacteria and viral infection. Traditional and novel approaches are used in drug discovery, which can be grouped into three categories [7]:

1. Random screening of a large number of compounds in search of desired biological properties.
2. Structural modifications of lead compounds, through the substitution, addition or elimination of chemical groups.
3. Rational drug design, including different approaches and techniques most of them with important computational component.

These approaches are not necessarily incompatible, and most companies try to use new methods to accelerate the discovery of new compounds. QSAR is a new technique based on the reasonable premise that the biological activity of compounds is a consequence of its molecular structure, provided we can identify those aspects of molecular structure that relevant to a particular biological activity.

QSAR is a part of chemometrics discipline that represents an attempt to correlate structural or property descriptors of compounds with activities. In other words, it is an indication of the explosion of techniques, procedures, and ideas, all relating in some fashion to attempt to summarise chemical and biological information in a form that allows one to generate and test hypotheses to facilitate an understanding of interactions between molecules. QSAR can also be referred to statistical analysis of potential relationships between chemical structure and biological activity.

The goal of structure activity relationship is to analyse and detect the determining factors for the measured activity for a particular system, in order to have an insight on the mechanism and behaviour of the studied system. For such purpose, the strategy is to generate mathematical models that correlate experimental measurements with a set of chemical descriptors determined from the molecular structure for a set of compounds.

The formulation of thousands of equations using QSAR methodology attest to a validation of its concepts and its utility in the elucidation of the mechanism of action of drugs at molecular level and more complete understanding of physicochemical phenomena such as hydrophobicity. It is now possible not only to develop models for a system but also to compare models from a biological database and to draw analogies with model from physical organic database.

1.3 History and Development of QSAR

More than a century ago, Crum-Brown and Fraser expressed the idea that the physiological action of a substance was a function of its chemical composition and constitution [8]. In 1863, Cros at the university of Strasbourg observed that toxicity of alcohols to mammals increased as the water solubility of alcohol decreased while in 1890's, Hans Horst Meyer of the university Marburg and Charles Ernerst Overton of the university of Zurich, working independently, noted that the toxicity of organic compounds depended on their lipophilicity [9]. Basing on biological experiments, they correlated partition coefficients with anesthetic potencies. Besides, Overton also determined the effect of functional groups in the increase or decrease of partition coefficients [10]. Afterwards, Lazarev in St. Petersburg continued where Overton and Meyer left off, applying partition coefficients to the development of industrial hygiene standards. Lazarev reported correlations on a log scale, and developed a system for estimating partition coefficients from chemical structure.

In 1893, Richet showed that the cytotoxicities of a diverse set of simple organic molecules were inversely related to their corresponding water solubilities and in 1939 the earliest mathematical formulation is attributed to Ferguson, who announced a principle for toxicity [8]. He observed the increase in anesthetic potency when ascending in a homologous series of either n-alkanes or alkanols to a point where a loss of potency, or at least no further increase occurred, using physical properties such as solubility in water, distribution between phases, capillarity and steam pressure.

Little additional development of QSAR occurred until the work of Louis Hammet (1937) within the field of organic chemistry, who observed that the addition of substituents to the aromatic ring of benzoic acid had an orderly and quantitative effect on the dissociation constant. He also correlated electronic properties of organic acid and bases with their equilibrium constants and reactivity. From empirical observation, he consequently derived the following linear relationship, the so called Hammet equation:

$$\log \frac{K}{K_0} = \rho\sigma \qquad 1.1$$

where the slope ρ is proportionality reaction constant pertaining to a given equilibrium that relates the effect of substituents on that equilibrium to the effect on the benzoic acid equilibrium. σ is a parameter that describes the electronic properties of aromatic substituents i.e. donating power. Based on Hammett's relationship, the electronic properties were utilized as the descriptors of structure [9].

Taft devised a way for separating polar, steric and resonance effects and introducing the first steric parameters, Es [11]. Working in the same direction, Swain studied the effects of field and resonance. He investigated the variation of reactivity of a given electrophilic substrate towards a series of nucleophilic reagents [10].

Free and Wilson partitioned the molecule in a different manner as Hammet. They postulated that the biological activity of a molecular set can be related with the addition of substituents, taking into account the number, type and position in the parent skeleton [10].

In 1962 Hansch and Muir published their brilliant study on the structure activity relationship of plant growth regulators and their dependency on Hammett constant and hydrophobicity. The parameter π, which is relative hydrophobicity of substituents, was defined in a manner analogous to the definition of sigma:

$$\pi_x = \log p_x - \log p_H \qquad 1.2$$

P_x and P_H represent the partition coefficients of derivative and the parent molecule, respectively. In 1964 Hansch and Fujita combined these hydrophobic constant with Hammett's electronic constants to yield the linear Hansch equation.

Hansch analysis is powerful technique for use in optimizing the activity of lead compounds. All physicochemical factors that relate to the transport and receptor interaction can be broken down into hydrophobic, electronic and steric component. Correlation between hydrophobic, electronic and steric components to biological activity can be summarized in an equation like below:

$$Log \frac{1}{C} = a\pi + b\sigma + cEs + d \qquad 1.3$$

where C is molar concentration of compounds, π, σ and Es is hydrophobic, electronic and steric component. a, b, c and d are regression coefficients. The combination of Hansch and Free-Wilson analysis in a mixed approach widens the applicability of both QSAR methods.

The linear free energy relationship (LFER) approach was contributed as the first attempt to predict the property of a compound from an analysis of its structure [12]. LFER methods are widely used for the development of quantitative models for energy-based properties such as partition coefficients, binding constants, or reaction rate constant. This is based on the pioneering work of Hammet, who introduced this method for the prediction of chemical reactivity. The basic assumption is that influence of a structural feature on the free energy change of a chemical process is constant for a co generic series of compounds. The basic LFER approach was later extended to the more general concept of fragmentation. Molecules are dissected into substructures and each substructure is seen to contribute a constant increment to the free energy based property. The promise of strict linearity does not hold true in most cases, so correction have to be applied in the majority of methods based on

fragmentation approach. Correction terms are often related to long range interaction such as resonance or steric effect.

Computer-assisted drug design (CADD), also called computer-assisted molecular design (CAMD), represent more recent application of computers as tools in the drug design process [9]. It is important to emphasize that computers cannot substitute for a clear understanding of the system being studied. A computer should therefore be considered as an additional tool to gain better insight into the chemistry and biology of the problem at hand. This tool has enabled the rapid synthesis of large number of molecules. Massive amount of data can be generated in relatively short period of time.

In the middle of the 20th century, two QSAR approaches now considered as classical were developed [7]:

1. Techniques based on the recognition of molecular features (fragments, groups or sites) and calculation, generally by regression analysis of the contribution that these patterns make to activity, assuming additively of the effects.
2. Techniques based on physicochemical parameter as structural descriptors. The rationale of this method is the fact that biological responses of the living organism to drugs are frequently controlled by lipophilicity, electronic and steric properties.

1.3.1 Data Set

Data set consists of compounds with molecular structure and biological activity; the compounds were divided between training and test set. Approximately 40 % were selected with a maximum dissimilarity algorithm and assigned to the test set, with the remaining 60 % assigned to training set [13]. The training set was used for QSAR model development and test set was used for model validation.

Other techniques that can be used to make a division of a data set into training and test set are based on sphere-exclusion algorithms [14]. The procedure

implemented in this method starts with the calculation of the distance matrix **D** between representative points in the descriptor space. Each probe sphere radius corresponds to one division into training set and prediction set. A sphere-exclusion algorithm consists of the following steps:

1. Select a compound with the highest activity.
2. Include this compound in the training set.
3. Construct a probe sphere around these compounds.
4. Include compounds, corresponding to representative points within this sphere, except for the sphere center, in the test set.
5. Exclude all points within this sphere from the initial set of compounds.

The procedure for division of a data set can also be done by sorting the list in increasing value of biological activity. Next, the odd numbered compounds are assigned to training set and even numbered compounds are assigned to prediction set or in the other way even numbered compounds are assigned to training set and odd numbered compounds are assigned in prediction set.

1.3.2 Descriptors

QSAR models are constructed by analyzing known or computed property data and series of descriptors representing the system characteristic. An important class of these descriptors belongs to the empirical parameter category derived from physical organic chemistry. These parameters focus on how chemical reaction rates depend on differences in molecular structure.

Encoding the molecules numerically allows an indirect link between structure and activity to be established. Descriptors are numerical quantities that characterize properties of molecules [11]; descriptors also can be defined as numerical values that encode certain aspects of molecular structure [12, 15]. For each structure in the data set, more than 200 descriptors can be calculated ranging from atom and bond counts to more detailed combinations of structural information. The relationship between biological activity and descriptors is:

Molecule activity $= f$ (molecule structure) $= f$ (descriptor) 1.4

The QSAR methodology begins with calculation of numerical descriptors for a set of compounds. Figure 1.1 shows how the generation of structural descriptors establishes the relationships between molecular structures and properties or biological activities.

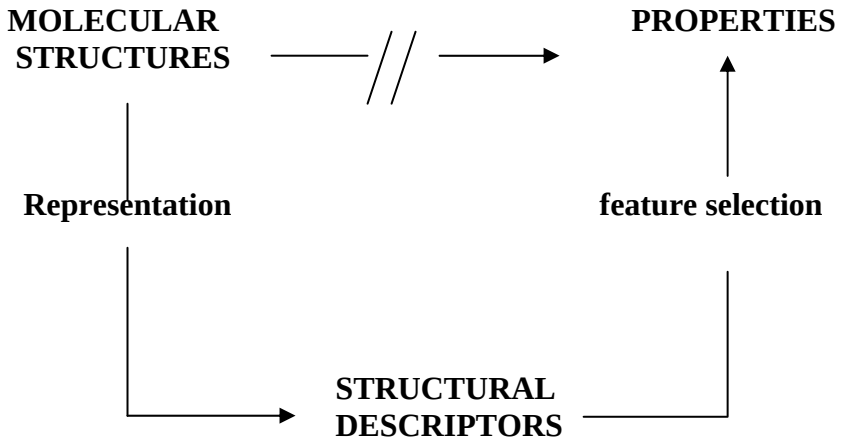

Figure 1.1: The general QSAR problem

Descriptors can be a quantitative property that depends on the structure of molecule. Various physicochemical parameters such as heat of formation, polarizability, hyperpolarizability, vibrational frequencies, etc have been used jointly with connectivity, topological indices and geometrical indices in order to get good model able to predict the anti bacterial activity [16].

The development of molecular structure descriptors is the most important part of any structure activity investigations because the descriptors must contain enough information to permit the correct classification of the compounds under study. Descriptors fall into three main categories: topological, electronic and geometric [17]. The following sections provide information and examples about each descriptor class to convey a clearer understanding of the descriptors routines.

1.3.2.1 Topological Descriptors

The structures of organic compounds can be represented as graphs. The theorems of graph theory can then be applied to generate graph invariants, which in the context of chemistry are called topological descriptors. The topological description of a molecule contains information on the atom-atom connectivity in the molecule, and encodes the size, shape, branching, heteroatom and the presence of multiple bonds [18, 19]. This graph description of molecules neglects information on bond lengths, bond angles and torsion angles, but is able to encode in numerical form the important atom connectivity information that determine a wide range of physical, chemical and biological properties. Topological indices are widely used as structural descriptors in quantitative structure-property relationships (QSPR) and QSAR models.

The Wiener index, W, defined in 1947, is widely used in QSAR and QSPR models as a part of topological descriptors, and it still represent an important source of inspiration for defining new topological indices. The path number W is defined as the sum of the distances between any two carbon atoms in the molecule, in terms of carbon-carbon bonds [17]. Hosoya extended the application of the wiener index by defining it from the distance matrix as the half sum of the diagonal elements of a distance matrix in the hydrogen depleted molecular graph [20].

Randic firstly introduced the concept of molecular connectivity in 1975. It is also called the connectivity index or branching index, to provide a topological index that could characterize the amount of branching in hydrocarbon molecules. This initial concept was extended by Kier and Hall to develop the well known χ indices [18]. They have found the branching index is seen to provide some basic information concerning the overall composition of the molecule.

Another example of topological descriptors is the electro topological states. This descriptor is a numerical value computed for each atom in a molecule, which encode information about both topological environments of the atom and the electronic interactions due to all other atoms in the molecule [12].

1.3.2.2 Electronic Descriptors

A large variety of electronic whole molecule descriptors have been used to encode the electronic features in QSAR investigations. The electronic environment of each molecule is estimated with the electronic descriptor routines. Electronic descriptors provide information about the overall charge distribution by calculating values such as the partial charges on each atom.

A number of electronic descriptors may encode the effects or strengths of intermolecular interactions. The more commonly recognized intermolecular forces arise from the following interactions; ion-ion, ion-dipole, dipole-dipole, etc. There are some examples of this descriptor, such as electric dipole moment, that encodes the strength of polar type interaction.

Molecular polarizability and molar refractivity are closely related properties that measure a molecule's susceptibility to becoming polarized. While descriptor related to intermolecular interactions are useful for predicting bulk physical properties and certain types of biological activities, they provide little direct information about the reactivities of compounds. This information is available through molecular orbital calculation [20].

The HOMO energy is roughly related to the ionization potential of a molecule, while the LUMO energy is related to the electron affinity. The magnitudes of these quantities are measures of the overall susceptibility of the molecule to losing a pair of the electron to an electrophile or accepting a pair of electrons from a nucleophile.

1.3.2.3 Geometric Descriptors

Biological activity is often related to the shape and size of the active compounds as well as the degree of complementarity of the compound and a receptor. With the given methods for generating three-dimensional molecular models of compounds, these models can be used to develop geometric descriptors.

Geometric descriptors capture information about the overall three-dimensional size and shape of molecules. As the name implies, they require that the molecules reside in accurate, three dimensional geometric conformations before descriptor generation. Examples of geometric descriptors: include the calculation of solvent accessible surface area and volumes and moment of inertia. These descriptors are useful in encoding steric effects that can occur between molecules.

Geometric descriptors appear frequently in QSAR of biological activity, especially when solvent accessible surface area information is used in conjunction with partial charge information to form the polar surface area descriptors. Surface area has a prominent effect on the interactions which occur between a drug molecule and its surroundings [20]. The other calculated descriptor for biological activity investigation is the molecular volume. The total molecular volume is taken as the sum of the contributions for each atom in the structure. The volume contributions of attached hydrogen atoms are also included in the final volume.

1.4 Feature Selection

Each descriptor contains useful information, but not all of these descriptors will be used to develop QSAR models. Feature selection was needed to reduce the number of descriptors. It is a step carried out in many analysis of reducing an initial too-large set of descriptors down to some smaller number that are felt to include the descriptors that matter [21].

The objective of feature selection is to identify the best subset of descriptors and to reduce the descriptors pool to a reasonable number; several stages of statistical testing are performed to remove descriptors that contain redundant information. Two methods to achieved feature selection:

i) Objective feature selection uses only the independent variable; the goal is to remove redundancy amongst the descriptors and to deter chance effects during model development. Pair wise correlations coefficient are calculated for all pairs

of descriptors, if r^2 value is greater than 0.8, one of the two descriptors will be rejected randomly.

ii) Subjective feature selection which also uses the dependent variable is used to identify the most information rich descriptor subsets which best map an accurate link between structure and a property of interest.

The genetic algorithm (GA) method can also be used to select the optimum number of descriptors for use in regression analysis. The GA could be useful technique for searching large probability space with a large number of descriptors for a small number of molecules. For example, this technique has been successfully applied to select the descriptors which can be used to correlate and predict effect concentration (EC_{50}) values of fluorovinyloxyacetamides compounds [22].

K-nearest–neighbor (KNN) analysis has also been used as variable selection procedure. In principle, this technique seeks to optimize simultaneously the selection of variables from the original pool of all molecular descriptors that are used to calculate similarities between compounds. KNN technique has been applied to select descriptors and establish the QSAR models for predicting the anticonvulsant activity of functionalized amino acid [23].

Searching all combination of descriptors is impractical so a logical approach is taken by combining an optimization routine. It has been shown to be very efficient in screening the reduced pool to identify optimal models [24]. Generalized simulated annealing (GSA) attempts to find models with the best configuration of descriptors that will produce low error for the training set compounds [25]. Once the initial model is evaluated for fitness, a perturbation is made by randomly replacing one (or more) descriptor with another from the reduced pool. If the new model is better than the first, the step is accepted and third model is produced via perturbation of the descriptors in the second model.

Multiple linear regression analysis can only handle data sets where the number of descriptors is smaller than the number of molecules, unless again a preselection of descriptors is carried out (e.g. by using GA). Genetic algorithm-multiple linear regression analysis (GA-MLRA) have been combined to make a new

classification and regression tool for predicting a compound's quantitative or categorical biological activity based on a quantitative description of the compound's molecular structure [26].

1.4.1 Genetic Algorithm (GA)

The GA approach is a general optimization method first developed by Holland [27] involves an iterative mutation/scoring/selection procedure on a constant-size population of individuals. The theory behind GA originates from the 'survival of the fittest' principle. Darwinian Theory states that individuals who possess dominant features will prevail in a population and produce children with even more superior features. In GA, models represent chromosomes while the descriptors comprising the model represent the genes encoding each chromosome. Mating and mutation allow GA to efficiently scan an error surface and assess the fitness for thousands of models.

The advantages of GA methods are: it searches the descriptor space efficiently and it can find models containing combination of descriptors or features that predict well as group but poorly individually [28, 29]. GA methods were used to select the optimum number of descriptors for use in regression analysis. The general GA scheme is shown in Figure 1.2.

1.5 Tools and Techniques of QSAR

QSAR studies include mathematical correlation between molecular structure and its activity. For quantitative modeling, two methods are primarily used to develop QSAR/QSPR models. Model complexity generally increases during the model development stages. Simple methods requiring low computational resources are examined first with the more complex and computationally demanding techniques being employed last in an effort to increase model quality.

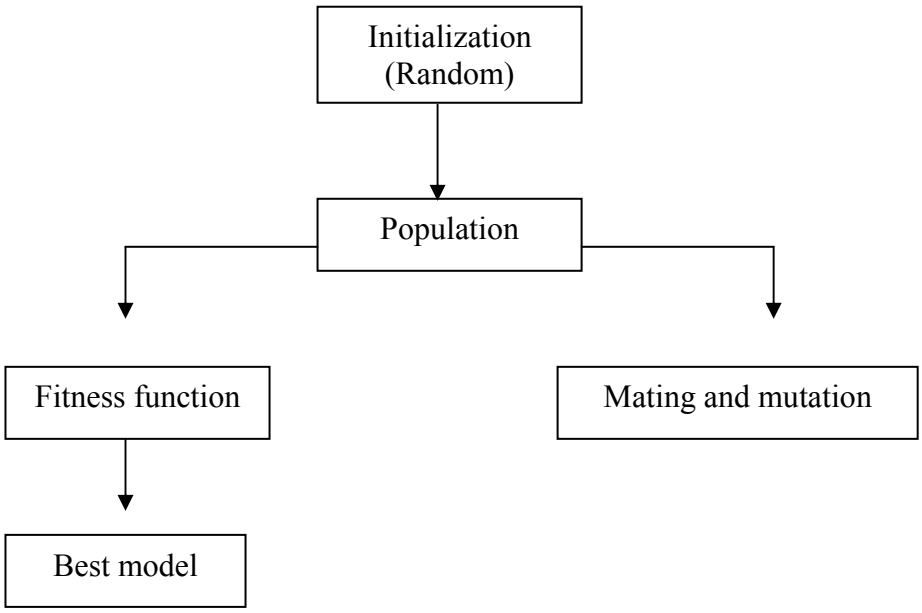

Figure 1.2: Flow diagram for the GA

The first and most widely used mathematical technique in QSAR analysis is multiple regression analysis (MRA). Regression analysis is a powerful means for establishing a correlation between independent variables which in this case usually include physicochemical parameter and dependent variable such as biological activity [22, 30].

1.5.1 Multiple Linear Regression Analysis (MLRA)

The goal of MLRA is to find the best subset of descriptors which provide accurate predictions for each compound in the training set. For each model, the values for descriptor coefficients and y-intercept are found that provide the most accurate mapping between input descriptors and property of interest. Generally, linear regression is represented by the equation below:

$$c = Rb \qquad 1.5$$

where c is matrix of molecular activity (n sample x 1), R is a matrix of descriptors (n sample x n descriptors) and b is model coefficients (n descriptors x 1).

Using the response matrix and the known activity of only one of the compounds c, the regression coefficients (equation 1.5) can be estimated as:

$$\hat{b} = (R^T R)^{-1} R^T c \qquad 1.6$$

where $\hat{b}$ is the regression vector, R^T is the pseudo-inverse of matrix descriptors, R is a matrix descriptors and c is activity of compounds.

$$\hat{c} = r_{unk}\hat{b} \qquad 1.7$$

(r_{unk}) is matrix of the known descriptor, it is possible to use the estimated regression vector (b) to predict the activity of unknown compounds ($\hat{c}$), by using this equation (equation 1.7).

Multiple regressions calculate an equation describing the relationship between a single dependent y variable and several explanatory X variables [31]. The independent variable, which in this case usually include the physicochemical parameter and biological data are assumed as dependent variable. The analysis derives an equation of the form [11]:

$$Y = a_1x_1 + a_2x_2 + a_3x_3 + \ldots..a_nx_n + e \qquad 1.8$$

The multiple correlation coefficient r^2 describes how closely the equation fits the data. If the regression equation describes the data perfectly then r^2 will be 1.0 [32, 33].

$$r^2 = \frac{SSR}{SST} \qquad 1.9$$

Where *SSR* is the explained Sum of Squares of *y* and *SST* is the total sum of the difference between the observed *y* values and their mean.

$$SST = \sum_{i=1}^{n} (y - \bar{y})^2 \qquad 1.10$$

SSR is the sum of the difference between the predicted *y* values ($\hat{y}$) and mean.

$$SSR = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2 \qquad 1.11$$

The major drawback of regression analysis is the danger of over fitting. This is the risk that an apparently good regression equation will be found, based on a chance numerical relationship between the *y* variable and one or more the *x* variable, rather than a genuine predictive relationship. When an over fitted model is used predictively, the predicted values for untested compounds will not be an accurate prediction of true values.

1.5.2 Partial Least Squares (PLS)

PLS was developed in the 1960's by Herman Wold as an econometric technique, but its most avid users are chemical engineers and chemometricians [33]. PLS has been applied to monitoring and controlling industrial processes; a large process can easily have hundreds of controllable variables and dozens of outputs.

PLS analysis calculates equations describing the relationship between one or more dependent variables and a group of explanatory variables [34]. PLS include two steps procedure; they are principal component analysis (PCA) and multivariate linear regression (MLR).

PLS analysis can be used in exactly the same way as regression, a single *y* (dependent) variable and two or more *x* (independent) variables are specified. PLS

always include all x variables in the analysis. As with regression, an equation is derived that allow the y values for unknown variables to be predicted from known x values [35]. Therefore, PLS is able to investigate complex structure activity problems, to analyze data in a more realistic way, and to interpret how molecular structure influences biological activity [10].

An important feature of the method is that usually a fewer factors (variables) are required. The precise number of extracted factors is usually chosen by some heuristic techniques based on the amount of residual variation. Another approach is to construct PLS model for a given number of factors on one set of data and then to test it on another, choosing the number extracted factors for which the total prediction error is minimized.

Recall the form of linear regression model is $c = Rb$ (equation 1.5) The difficulty often encountered when solving for b is that the $R^T R$ matrix is not invertible because of redundancy in the variables. Principal component regression (PCR) eliminates this redundancy by constructing a new matrix U with column that is linear combinations of the original columns in R. Using the U matrix, a new model is written as shown in equation 1.12:

$$c = U\tilde{b} \qquad 1.12$$

The technique of PLS is similar to PCR with the crucial difference that the quantities calculated are chosen to explain not only the variation in the independent (X) variable but also the variation in the dependent (Y) variables as well. PCR produces the weight matrix reflecting only the covariance of the predictor variables, while PLS regression includes the response variable Y in the process of reduction of the variables, and so the covariance is between the independent and dependent variables.

PCR and PLS use different approaches for choosing the linear combinations of variables for the columns of U. PCR only uses the R matrix to determine the linear combinations of variables but in PLS technique, the covariance of the

measurements with the concentrations is used in addition to the variance in *R* to generate *U* [36]. The illustration of the difference between PCR and PLS is shown in Figure 1.3.

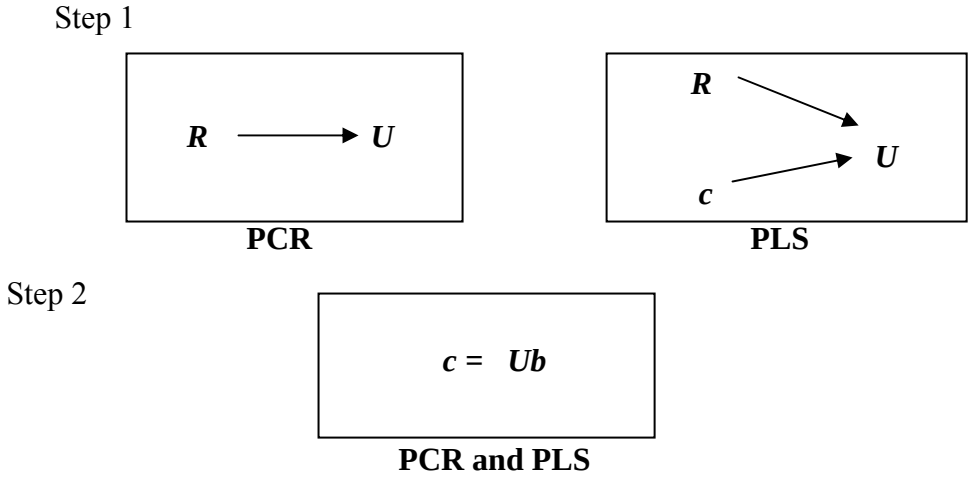

Figure 1.3: Illustration of the difference between PCR and PLS

U is the score matrix from PCA, which defines the location of the samples relative to one another in row space. The score matrix is related to the original matrix *R* (matrix of descriptor) in the following manner:

$$R = USV^T \qquad 1.13$$

Where *U* is the score matrix, *V* is a matrix containing the loadings and *S* is a diagonal matrix. The orthonormal property of *V* (i.e., $V^TV = I$) can be used to solve equation 1.13 for *U* as follows:

$$U = RVS^{-1} \qquad 1.14$$

The following equation is possible to solve equation 1.12 and can be used to predict the activity of unknown compounds:

$$\hat{\tilde{b}} = U^T c \qquad 1.15$$

where c is matrix of activity, U^T is pseudo-inverse of score matrix from PCA and $\hat{\tilde{b}}$ is regression vector.

1.6 Applications of QSAR

The major goal of QSAR in chemical research is to predict the behavior of new molecules, using relationships derived from analysis of the properties of previously tested molecules. QSAR studies represent one of the best methodologies in computer based drug design, offering valuable information about biological activity and providing a computationally inexpensive methodology to design of potential bioactive drugs.

MRA was used to generate the QSAR models. These models were constructed by correlating the topological descriptors and anti tumor activity of 20 (S)-campotechin derivatives. Good QSAR models can be used to instruct the designing and predicting the anti tumor activity of new analogues [37].

QSAR approach also can be used to search new agents against *Mycobacterium tuberculosis* (*M. tuberculosis)* and other typical mycobacteria. It is significant due to the lack of effectiveness of known anti tuberculosis agents against opportunistic pathogens as a consequence of rapidly emerging resistance [38].

QSAR method was employed by using the hydrophobicity and electrophilicity as parameter to investigate the structural features that affect the toxicity of nitrobenzene derivatives to yeast and response-surface analysis was performed to develop a robust QSAR for predictive use [39]. QSAR model also have been developed between hydrazide potencies against *Escherichia coli* (*E. coli*) and *Sacharomyices cerevisae*. The study shows that an extra thermodynamic relationship can be established between two different cell systems [40].

The inflammatory process is necessary for survival against pathogens and injury, but sometimes the inflammatory response is aggravated and sustained without benefit. A large number of Homoisoflavanones have been isolated several genera within the *Hyacinthaceae* family and have anti-inflammatory properties. The biological data was then correlated to the physicochemical descriptors of the compounds by applying statistical regression analysis and also to establish a quantitative structure activity relationship model with reliable predictive ability as the potential degree of anti-inflammatory activity of compounds within this class [41].

QSAR studies are being applied in Environment assessment; toxicity to aquatic life form is one of the crucial factors in evaluating the environmental risks of man-made chemicals. Chemicals could jointly cause toxic effects to fish at concentration as low as 2% of their individual inhibition concentration (IC_{50}). The application of QSAR models derived from single chemicals toxicity assay are used to predict concentration of component in mixtures that would jointly cause 50% inhibition of microbial respiration [42].

Chemical and biological transformations, and degradations, play a role in the transport and mobility of such chemicals in the environment. Volatile Organic compounds (VOCs) are a class of organic chemicals largely present in the troposphere because of their vapor pressure. By using QSAR modeling can be used to predict the rate constant for hydroxyl radical trophospheric degradation of 46 heterogeneous organic compounds [43]. A variety of QSAR paradigms have been presented as possible computational tools to aid with the rapid assessment of endocrine disruptions potential for environmentally relevant component [44].

A large number of environmental chemicals known as endocrine-disrupting chemicals (EDCs) are suspected of disrupting endocrine functions by mimicking or antagonizing natural hormones in experimental animals, wildlife, and humans. EDCs may exert adverse effect through a variety of mechanisms, including estrogen receptor (ER)-mediated mechanisms of toxicity. Consequently, more than 58,000 environmental and industrial chemicals have been identified as candidates for possible experimental testing. QSAR could be used as an inexpensive prescreening

tool to prioritize the chemicals for further testing and to classify of chemicals according to their ability to bind the estrogen receptor [45].

The new approach of QSAR models is by developing a drug discovery strategy that employs QSAR models for chemical database mining. The approach classified the lead molecules to active and inactive classes also to predict their biological activity [12].

1.7 Overview of Multidrug Resistance *Mycobacterium tuberculosis*

TB, or tuberculosis, is a disease caused by bacteria called *Mycobacterium tuberculosis* (*M. tuberculosis*). It can affect several organs of human body, including the brain, the kidney and the bones; but most commonly it affects the lungs (pulmonary tuberculosis). It is estimated that one-third of the world's population is infected with this bacteria. While only a small percentage of infected individuals will develop clinical tuberculosis, each year there are approximately eight million new cases and two million deaths. *M. tuberculosis* is thus responsible for more human mortality than any other single bacterial species [46].

Since tuberculosis spreads easily when people are in close contact with an infected person, it was more common in towns than in the countryside. People often came to towns to trade their goods or do other business. Between sixteenth and nineteenth centuries, many of the new arrivals in the major cities of Europe were consumed by tuberculosis plus other infectious diseases. A city's population was maintained only by a steady supply of healthy young people coming to make their fortunes. The Industrial Revolution, which began in the late seventeenth century in England and perhaps a hundred years later in the United States, brought more people into the urban areas and city life became more perilous. People could not escape the risk of tuberculosis infection even in their own homes, away from the factories. All these factors created the perfect breeding ground for tuberculosis which became an epidemic in Europe and later in the United States.

A number of efficacious anti tuberculosis agents were discovered in the late 1940's and 1950's with the last, rifampin introduced in the 1960's [47]. These agents have reasonable efficacy and when used in combination should preclude the development of drug resistance. However in 1962, Eleanor Roosevelt died of tuberculosis. It was learned that *M. tuberculosis* was resistant to this agent. The use (or in most cases misuse) of these drugs has lead over the year to an increasing prevalence of multi-drug resistant (MDR) strains and there is now an urgent need to develop new effective agents.

1.7.1 *Mycobacterium tuberculosis*

The genus *Mycobacterium* consists of gram positive bacilli with distinctive cell wall characterized by the presence of unusual glycolipids. A number of mycobacteria are pathogenic for man but the most important is undoubtedly *M. tuberculosis*, the causative agents of tuberculosis [47].

M. tuberculosis is a part of tubercle bacilli species, it grow well (eugenic) on egg media containing glycerol or pyruvate. Colonies resemble breadcrumbs and are cream colored. Films show clumping and cord formation especially on moist medium and it is usually resistant to thiophene 2 carboxylic acid hyrazide (TCH) is nitrase positive, aerobic and susceptible to pyrazinamide.

M. tuberculosis is an obligate aerobe which grows at different rates within cavities, caseous foci, and macrophages. The doubling time is 12-14 hours as compared to 0.25-1 hour for most other pathogenic bacteria. Since the efficiency of many bacterial agents is directly proportional to the rate of growth, eradication of infection requires prolonged therapy (6-18 months).

1.7.2 How Does Tuberculosis Spread

The TB germ is carried on droplets in the air, and can enter the body through the airway. A person with active pulmonary tuberculosis can spread the disease by coughing or sneezing. To become infected, a person has to come in close contact with another person having active tuberculosis.

The process of catching tuberculosis involves two stages: the first stage of the infection usually last for several months. During this period, the body's natural defenses (immune system) resist the disease, and most or all of the bacteria are walled in by a fibrous capsule that develops around the area. Before the initial attack is over, a few bacteria may escape into the bloodstream and be carried elsewhere in the body, where they are again walled. In many cases, the disease never develops beyond this stage, and is referred to as TB infection. If the immune systems fails to stop the infection and it is left untreated, the disease progress to the second stage, active disease. There, the germ multiples rapidly and destroys the tissues of the lungs (or the other affected organ). Sometimes, the latent period is many years, and the bacteria become active when the opportunity presents itself, especially when immunity is low.

The second stage of the disease is manifested by destruction or consumptions of the tissues of the affected organ. When the lungs is affected, it results is diminished respiratory capacity, associated with other organs are affected, even if treated adequately, it may leave permanent, disabling scar tissue [48].

Usually, the initial diagnostic/screening test for tuberculosis is the skin test. A small amount of fluid is injected under the skin of the forearm; the fluid contains a protein derived from the microorganism causing TB, and is absolutely harmless to the body. The area is visually examined by a health professional after 48-72 hours to determine the result of the test.

1.8 Minimum Inhibition Concentration (MIC)

QSAR have been used widely to predict the hazard of untested chemicals with already tested chemicals by developing statistical relationship between molecular structure descriptors and biological activity [49]. The principles of determining the affectivity of noxious agents to a bacterium were well enumerated at the turn of the century, the discovery of antibiotic made these tests (or their modification) too cumbersome for the large numbers of test necessary to be put up as a routine analysis.

Diffusion test widely used to determine the susceptibility of organisms isolated from clinical specimens have their limitation; when equivocal results are obtained or in prolonged serious infection e.g. bacterial endocarditis, the quantitation of antibiotic action needs to be more precise. The way to a precise assessment is to determine the MIC of the antibiotic to the organism concerned.

MIC is the lowest concentration of the antibiotic which will inhibit the growth of microbes. Dilution methods are used to determine the MIC of antimicrobial agents [50]. In dilution test, microorganisms are tested for their ability to produce visible growth on a series of agar plates (agar dilution) or in microplate wells of broth (broth microdilution) containing dilution of the antimicrobial agent. The lowest concentration of an antimicrobial agent which will inhibit the visible growth of a microorganism is known as the MIC

MIC methods are widely used in the comparative testing of new agents. In clinical laboratories they are used to establish the susceptibility of organisms that give equivocal result in disk test, for test on organisms where disk test may be unreliable, and when a more accurate result is required for clinical management.

1.8.1 *Escherichia coli*

The bacteria *E. coli* was named after the Austrian doctor, Theodor von Esherich (1857-1911), who first isolated the genus of bacteria belonging to the family enterobacteriaceae, tribe Escherichia. This bacterium is the common inhabitant of the intestinal tract of man and other animal, it is needed to breakdown cellulose and it assists in the absorption of vitamin K, the blood clotting vitamin [51].

E. coli is a motile species, it can produce acid and gas from lactose at 44^{o}C and at lower temperatures, is indole positive at 37^{o}C, MR positive, fails to grow in citrate and is malonate and gluconate negative. It is H_2S negative and usually decarboxylates lysine [52]. The structure of *E. coli* [53] is shown in Figure 1.4:

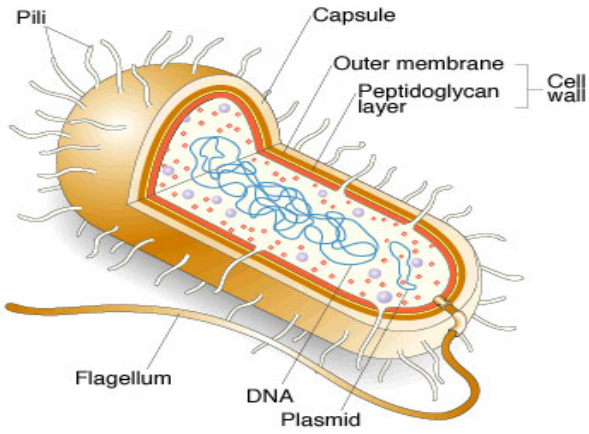

Figure 1.4: Structure of *E. coli*

E. coli is one of the normal bacterial floras of the gastrointestinal tract of poultry. Ten to fifteen percent of the intestinal coliforms in chicken are of pathogenic serotypes. Colibacillosis is a common systemic infection caused by *E. coli* in poultry. The disease causes considerable economic damage to poultry production world wide. As least four types of *E. coli* cause gastrointestinal disease in human: enter pathogenic (EPEC), enterotoxigenic (ETEC), enteroinvasive (EIEC) and verotoxigenic (VTEC).

The EPEC strains have been associated with outbreaks of infantile diarrhea and identified serologically. ETEC strains are thought to cause gastroenteritis in both adults and children. While EIEC strains cause diarrhea to that in shigellosis. The strains associated with invasive enteric infections are less reactive than typical *E. coli* and VTEC strains derive their name from their cytotoxicities on Vero cells in tissue culture. They have been associated with haemolytic uraemic syndrome and haemorrhagic colitis.

1.9 Database Mining

Pharmaceutical lead compounds traditionally have been discovered by isolation of natural products from fermentation broth and plants extracts, and by screening corporate chemical databases. Recently, this process has been assisted by structure based rational drug design technology. Drug design is one of the most important fields of study for bioinformatics. A major goal of drug design is to discover and optimize novel chemical substances that specifically interact with target molecules and, as a consequence, compensate or reverse disease process [54].

Drug abuse continues to remain one of the most difficult and a costly issue of modern society and cocaine is among the most heavily abused and devastating illicit substances. QSAR models have been developed to correlate structural features of the dopamine transporter (DAT) ligands and their biological activities. It also have been employed to search new lead compounds in the national cancer institute (NCI) database and yielded five compounds that are suitable for development as novel DAT inhibitors [55].

Another application is in the search for anticonvulsant agents to treat epilepsy. Epilepsy is a chronic disorder, characterized by recurrent unprovoked seizures. Currently, the main treatment for epileptic disorder is the long term and consistent administration of anticonvulsant drugs. Therapies have failed to adequately control this disorder, documenting the need for new agents with different mechanisms of action. Development of variable selection KNN QSAR models have

been used to mine external chemicals databases or virtual libraries for lead identification. This strategy was successfully applied for the discovery of novel anticonvulsant agents [56, 57].

The national cancer institute (NCI) USA has been carrying out invitro screening of compounds to determine their in vitro inhibitory activity of cell growth in the NCI 60 human cancer cell lines for the purpose of anticancer drug discovery. These Web-based data mining tools allow robust analysis of the correlation between the in vitro anticancer activity of the drugs in the NCI anticancer database, the protein levels and mRNA levels of molecular targets (genes) in the NCI 60 human cancer cell lines for identification of potential lead compounds for specific molecular targets and for study of the molecular mechanism of actions for a drug molecule [58].

1.10 Research Scope

This study focused on developing QSAR models that correlate biological activity (e.g., anti bacterial, anti tuberculosis) and chemical structures. The validated QSAR models were applied to mining chemicals in a large database. Database mining is one of the most important follow-up applications of QSAR model development. The proposed model can be utilized to select compounds that have similar structural attributes as the active compounds in the training set and they are expected to demonstrate anti bacterial and anti tuberculosis activity. The compounds used in data set were limited to those that have been extracted from natural products, while the second data set consists of compounds derived from plant terpenoids.

1.11 Research Objectives

The main objectives of this research were:

1. To develop computer models that correlate biological activity of chemical compounds in natural products with their structural characteristics.

2. To apply of the QSAR models in screening a large library of compounds (database mining).

1. 12 Significance of Research

One potential contribution of this research is in the utilization of our natural products to develop anti bacterial agents. Successful development of new agents will undoubtedly increase the value of our natural resources. Terpenoids are also a class of compounds that have been extracted from natural products; they can be used to combat the growth of *M. tuberculosis* bacteria. As discussed previously there is an urgent need to develop new effective agents against *M. tuberculosis*. Billions of dollars are spent each year by the drug and chemical companies of the world in the effort to study the relationships between molecular structure and its bioactivities for generating new drugs. By using QSAR models, we can correlate structural feature of the compounds isolated from plants with their biological activity and the models can be used to predict the activity of new compounds. Knowledge from this study can be used in the production of new drugs in pharmaceutical industries.

1.13 Layout of the Thesis

This thesis is organized into five chapters. Chapter 1 describes the background of research, some review of the literature to understand the issues and formulate research problem. The review describes about QSAR, GA, tools and techniques of QSAR, overview of multidrug resistance *M. tuberculosis* and MIC. It also presents the research scope, research objective, significance of research, and layout of the thesis.

Chapter 2 presents the research methodology employed in conducting the study. Two main approaches were adopted; development of QSAR models and application of these models in database mining. QSAR models will be used to

predict the biological activity of unknown compounds not included in QSAR model development (prediction set) and it also can be further exploit for the design and discovery of new potent anti bacterial agents. Database mining can be used to search compounds that have similar attributes as the active compounds in the training set.

Chapter 3 presents the results from data set which consists of compounds isolated from natural products with biological activity against *E. coli*. It describes development of QSAR models to predict the anti bacterial activity of unknown compounds, followed by application to search for new compounds in database mining. Results of biological testing of selected compounds are also presented.

Chapter 4 presents the results of study in which plant terpenoids against *M. tuberculosis* was used as data set. It explains about the QSAR model development and it's validation by predicting the anti tuberculosis activity of compounds not included in the model development process. These models were later used to search new potent anti tuberculosis agents. Finally, the activities of the new lead molecules were tested experimentally by using disk diffusion method.

Chapter 5 presents the conclusions of this study. The report culminates with some suggestions for future research.